Statistical Inference Course Project - Part 1

Author	Date	GitHub
Oszkar Jozsa	February, 2015	[link]

Overview

Description

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

- Show the sample mean and compare it to the theoretical mean of the distribution.
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
- Show that the distribution is approximately normal.

Solution

I will save a simple data set of 1000 simulations into a matrix and compare it to a theoretical normal distribution using some plots (base and ggplot) and simple calculations.

Solution

Simulations

First, we set up the parameters `number`, `lambda` and `n`. Then we'll set the seed for reproducibility. (also, ggplot is loaded here as we'll need it later)

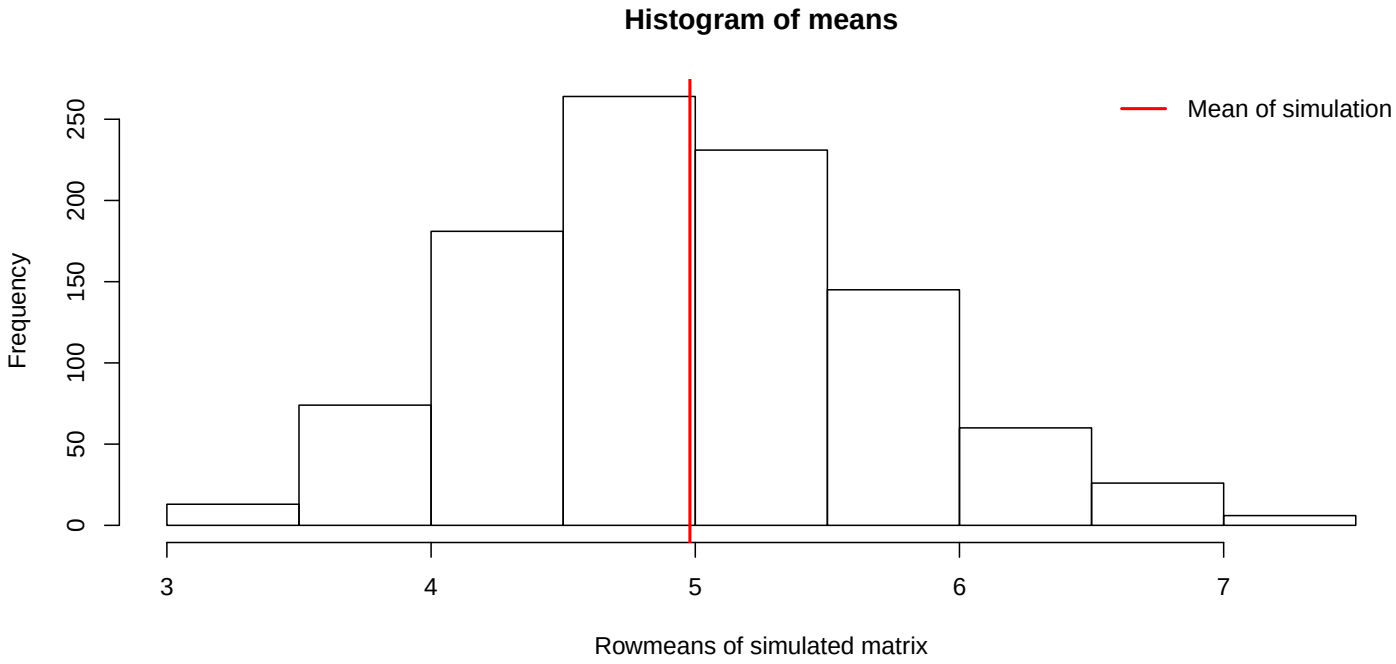
```
library(ggplot2)

number <- 1000
lambda <- 0.2
n <- 40

set.seed(543)
```

Next, the simulation is done with the parameters above and saved into `simulation` (1000-by-40 matrix). We also calculate the rowmeans here.

```
simulation <- matrix(rexp(number * n, rate=lambda), number, n)
means <- rowMeans(simulation)
```

We create a simple histogram to visualize the result of the simulation. Also, a red line is added for displaying the mean of the simulation.

```
hist(means, xlab = "Rowmeans of simulated matrix")
abline(v = mean(means), col = "red", lwd = 2)
legend(x = "topright", legend = c("Mean of simulation"), col = c("red"), bty = "n", lwd = 2)
```



Sample Mean versus Theoretical Mean

```
mean <- mean(means)
theoretical_mean <- 1 / lambda
mean
```

```
## [1] 4.979703
```

```
theoretical_mean
```

```
## [1] 5
```

Simulated (measured) mean is very close to the theoreticcal mean. We could get even closer by increasing `number` (more simulation).

Sample Variance versus Theoretical Variance

```
variance <- var(means)
theoretical_variance <- (1 / lambda) ^ 2 / n
variance
```

```
## [1] 0.5574497
```

```
theoretical_variance
```

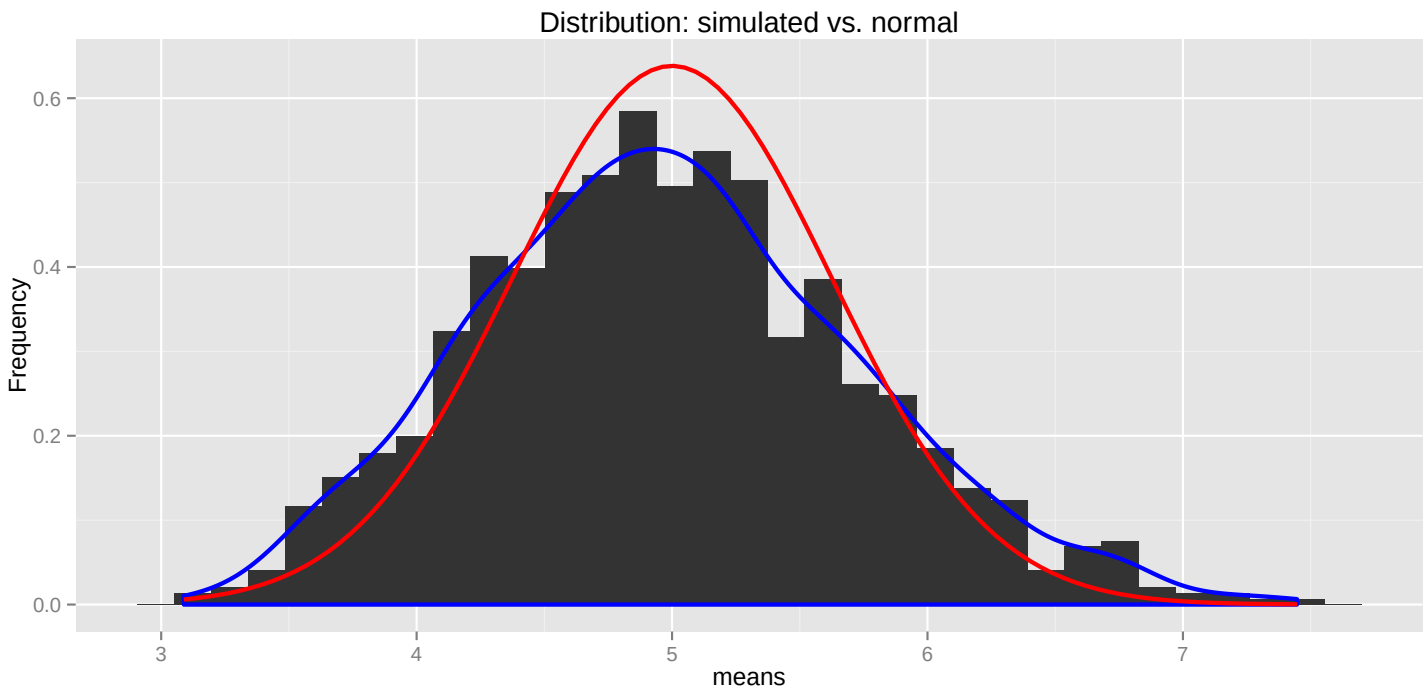
```
## [1] 0.625
```

Variance differs a bit more than the mean did, but still close to the theoretical one.

Distribution

The histogram below shows how close the distribution is to normal. Blue line is a density estimation of our measured data. Red line is a true normal distribution.

```
ggplot(data.frame(means), aes(x = means)) +
      geom_histogram(aes(y=..density..)) + geom_density(colour = "blue", size = 1) +
      stat_function(fun = dnorm, args = list(mean = theoretical_mean, sd = theoretical_variance), color = "red"
      labs(y = "Frequency") + labs(title = "Distribution: simulated vs. normal")
```



Lastly, a QQ plot to compare measured and theoretical distribution (see this article). Lines are almost exacly lined up which shows us that the measured is very close to the theoretical one. A 45 degree `abline` represents the theoretical QQ line.

```
qqnorm(scale(means)); qqline(scale(means), col = "blue"); abline(0, 1, col = "red")
legend(x = "topleft", legend = c("Simulation", "Theoretical"), col = c("blue", "red"), bty = "n", lwd = 2)
```